VALIDATION, EDITING AND EXPANSION IN A DEREGULATED ENVIRONMENT

by Joseph S. Lopes
Applied Energy Group, Inc.

INTRODUCTION

Load Research has had a long and varied history, having been carried on the coattails of a variety of applications since the early days of the 1960's and 1970's. With the collection of load data associated with load research came the need for quality assessment, decisions on how (and whether) to fix problem data and, once data was sufficiently clean, expanding results to the population, whether the load data represented one customer or part of a group representing tens of thousands.

Whoever coined the phrase “garbage in, garbage out” was certainly thinking of situations encountered in load research. With the high cost of load survey equipment and operations, small carefully crafted samples meant that quality from each site was paramount. Consider that one sample site might represent the pattern of tens of thousands of customers and you can appreciate the implications of including problem data in an expansion of results to model population load characteristics.

Problem data and data loss are caused by a combination of factors, including equipment failure, data communications losses, human error, neglect, weather, computer failures and bad luck. Equipment and procedures have improved over time, but data problems have not disappeared. Those of us who have been in the industry for many years can look back on the many hours spent fixing data problems, only to have technology eliminate them, or so we think - and create new ones.

In the early days of load research, between approximately 7,000 to 10,000 BC (10,000 days Before Competition, or between 1968 and 1975), magnetic tape recorders with bulky cartridges, and later cassettes, were used for load studies. This was a major improvement over strip charts, which were hard to read and generally unreliable, somewhat like stone tablets. Mag tapes did not solve all the problems. The primary data problems were caused by tapes jamming or outages, which caused missing or too few intervals of data, and bad splices, which caused mismatches between the end of one tape and the beginning of the next. When the customer lost power, the recorder stopped, too. In about 7,000 BC (1975), man discovered batteries! This enabled a time pulse to be continuous on recorders, so if the customer lost power, the recorder did not. This eliminated the uncertainty about when the interruption occurred, particularly when there were several. Then, man discovered electronic recorders about 3,000 BC (mid 1980's), and this virtually eliminated the mechanical failures associated with mag tape (Cro-Mag) recorders. Further advancements in electronics and means to “download” data and install equipment were also made in the past 15 years, such as an optical probes, telephone modems, wireless communications and meters “under the glass”.

Over those years, the reasons for collecting data changed, from PURPA[1] compliance, to conservation studies, load management, rate design, cost-of-service studies, demand-side management, technology assessment, billing, profitability, competitive threats and, most recently, load profiling to enable reconciliation of sales from multiple suppliers within a service area. The issues of data quality have remained, with only slight differences, and the turnover of experienced staff has meant that some of the experience and techniques for processing and analyzing load data has been lost.

[1] Public Utility Regulatory Policy Act of 1978, which required all utilities to collect load research data on major rate classes comprising more than 10% of retail sales.

As a result, this paper is being written to refresh and reminisce for the old timers, teach the new generation, and highlight the similarities and differences associated with data validation and editing under the new deregulated utility environment. As we approach 2000 AD (that's about 2,000 days After Deregulation, which began about 5 years ago), we need to take a fresh look at validation and editing techniques and their implications on the new applications for the data.

WHY VALIDATE AND EDIT?

Because of the relatively high cost and (sometimes) customer inconvenience associated with installing, maintaining, and processing load data from customer sites, the number of recorders that can be installed is limited by budget, priorities, manpower and other resources. Advances in sample design and standards for precision have meant that only a few hundred samples are considered sufficient to provide for the objectives of load studies and satisfy the requirements of both internal and external (including regulatory) criteria. However, this has meant that there is little margin for error and data must be recovered from a high percentage of the sites for which a significant investment in equipment and other resources has been made. A typical residential load study, for example, involves only 100 - 200 sample points, with careful segmentation to optimize the sample design. Thus, every sample data point lost means 0.5% to 1% of the data is lost - so a high data quality rate is essential.

The main objective of data validation is to identify data intervals that are either not accurate, missing or not representative. Recorder or data communications failures will often result in missing or zero intervals because the data was not properly recorded and/or transmitted. These problems must be identified to prevent skewing of the resulting class load profile expansion, since applying "bad" data can mean multiplying zeros, rather than valid energy totals, by up to tens of thousands, leaving "holes" in the data and a drop in load not reflective of the real situation.

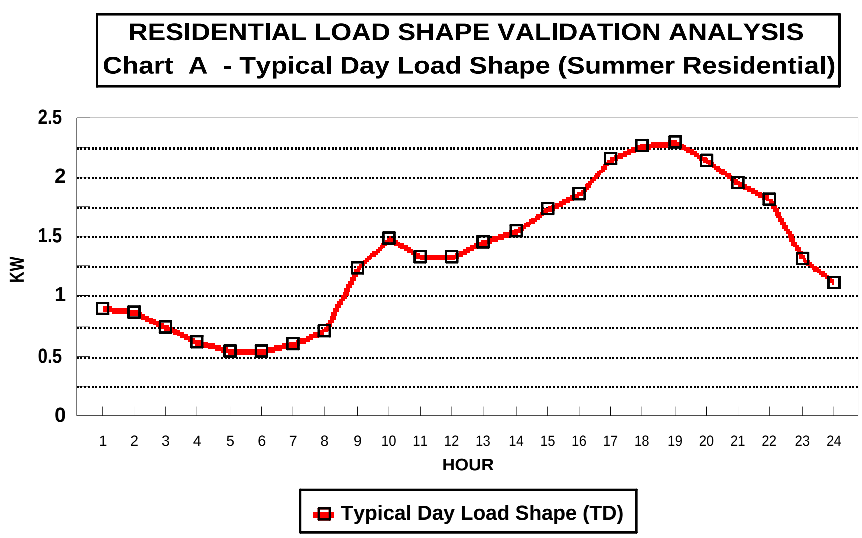
Outages at the customer are the next major problem. While these may be recorded accurately as zero use by recorders with battery backup, it is still advisable to patch the data to prevent these presumably isolated case from being applied to the hundreds, thousands or tens of thousands in the population that the sample site represents. Typically, an outage will result in a drop or reduction in load when measuring intervals (typically 15 minute or 60 minute) of any time length. Outages of more than two intervals are easy to visually identify because the load will often drop to zero. Once power is restored to the customer, the interval data will typically reflect a spike of higher usage, often referred to as "payback", where appliances, particularly heating and cooling, as well as water heating, refrigeration and other loads, must compensate for their cycling patterns being disturbed and must also recover the air/water temperatures their thermostats require. The drop or zero use, as well as the post-outage payback, are not representative of the customer's uninterrupted load, so editing must be applied to both the outage period and the payback period.

Thus, the risks of improper (or non-existent) data validation are great, and the design of effective validation tests is essential to reducing this risk. Once the problem intervals are identified, the decision must be made as to whether to "patch" the data so it can be included or to exclude the data altogether. The risk of exclusion of the data is that too much data loss can erode the sample sizes and resulting precision. Whether to exclude or patch and how to patch is often a judgment call by the analyst, and editing rules are important to maintain the integrity of the data. Too little editing of problem data is akin to ignoring a problem, which can be catastrophic if bad data is allowed to skew the load shapes. Too much editing can also be ill-advised, since the integrity of the data is important and "good" data should not be replaced just because it may not pass arbitrary validation tests. Load shape data is very variant and this natural variability of the data should not be eliminated by too much smoothing or averaging of the data. Patching of outages, for example, should be designed to restore the outage and following payback period to values that would likely have occurred if there was not an outage given the weather, day type, season and other conditions of the period involved.

VALIDATION TESTING CRITERIA

Validation tests are typically designed to flag data for later editing by analysts, using criteria designed to identify intervals with problems as "bad" and those without problems as "good". As with many statistical analysis functions, there are two types of error associated with these validation flags. Type I error is where data is rejected under a set of criteria when it should be accepted, and Type II error is where data is accepted when it should be rejected. Ideally, all bad data would be rejected and all good data would be accepted, but the effectiveness of the criteria will be the critical factor in properly flagging the data and minimizing both Type I and Type II error. In terms of load data, Type II error (accept bad data) is more serious because of the implications of including bad data intervals in an expansion, as mentioned already. Type I error (rejecting good data) has less risk, primarily from depleting the sample complement and diminishing precision.

The following discussion, highlighted by graphic illustrations, will describe data validation criteria using a typical summer day for a residential customer and several problem days where outages occurred. The key to effective validation is to identify those intervals - and only those intervals - that are a problem. Editing rules would then be applied to determine what to do, if anything, about these problems. Effective and efficient validation tests flag only a small number of intervals, and capture all the serious problems, minimizing Type II error (accept when it should reject), but also minimizing Type I error (reject when it should accept). A rule of thumb for typical data collected from current vintage equipment is that no more 2% of data is flagged, with less than 0.5% actually needing to be fixed



(patched or excluded).
In Chart A above, a typical residential load shape for a summer day is shown, with its distinct pattern of

morning activity (lights, cooking, water heating), cooling load in the afternoon, followed by evening activity (lights, cooking, TV, some cooling). If there was an outage, however, the load shape might look like the following:

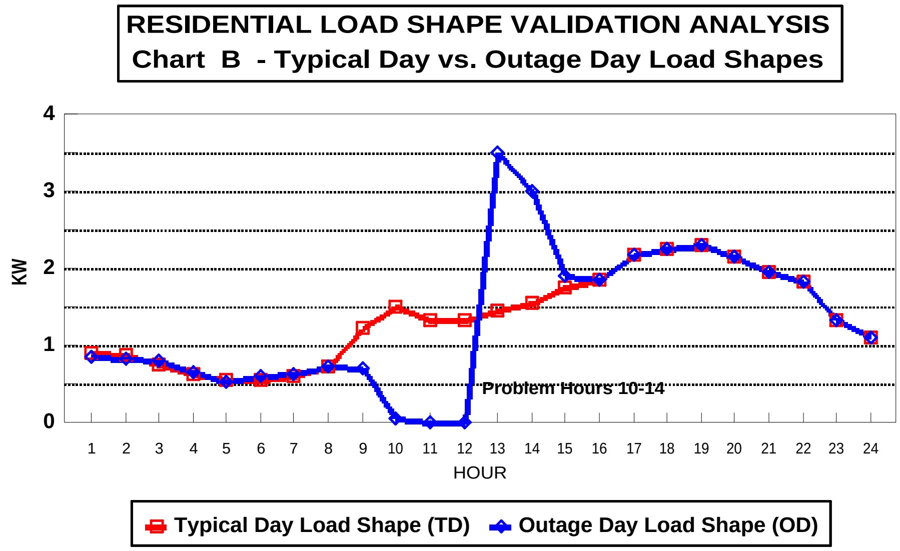


Chart B illustrates the effect of an outage (overlaid on the typical load shape) that began before hour 9 and continued until Noon, then restoral of service, which led to a spike and payback over several hours until about hour 15 (3 pm), when loads were back to normal.

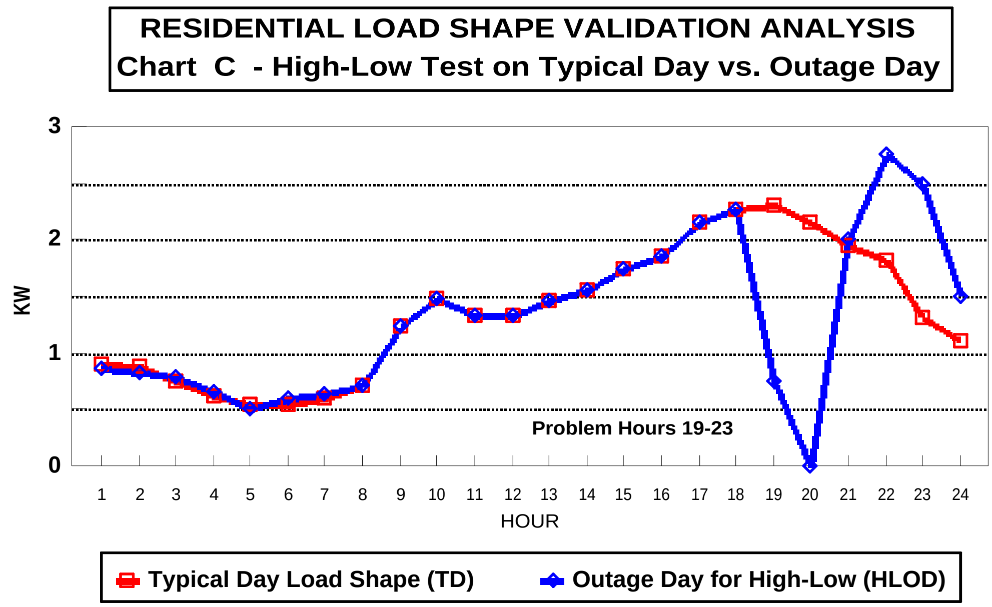
Through validation criteria, the problem hours of 9 through 15 should be identifiable. Certainly, a visual inspection of the data would alert the analyst to the problem. However, with the amount of data, and the increasing need for near real-time turnaround, automated routines are needed. The most common of these are listed below:

<u>High-Low Test</u>

This test is routinely applied for billing determinants, since a history on each customer can be used to flag unusual usage levels. Similarly, this test can also be applied to interval load data. As indicated in Chart C, however, depending on when the outage occurs, it may not create a significantly higher peak, so may not be identifiable as a problem interval.

Chart C below indicates a typical outage, overlaid on a typical load shape. The outage began around hour 19 and continued to after hour 20, followed by a spike and nearly recovery by hour 24. While this is clearly problem data, a high-low test may only flag the "0" value that would occur when an outage is long enough to cover a full interval - not always the case, and the payback spike is hardly higher than the afternoon peak because it occurs during a normally decreasing period late at night. In addition, if this day was not during the seasonal peak, the spike would likely be lower than a normally-occurring peak during a more extreme weather day, defeating the effort of the validation criteria to flag problem data.

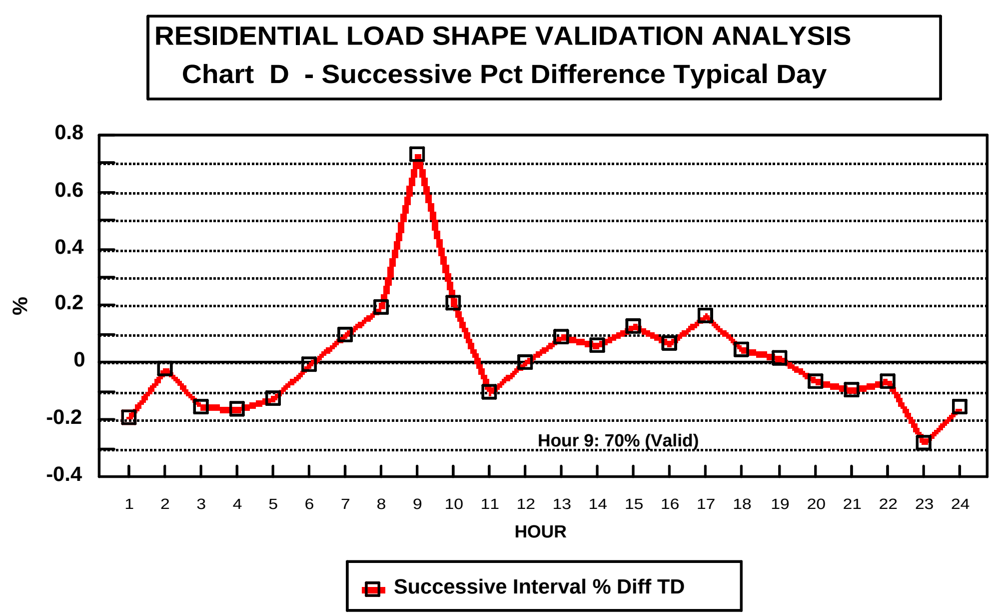
Therefore, while a High-Low test is a valid and generally successful one for monthly consumption and even billing demand, it could not be expected to meet the objectives of an efficient and effective validation test.

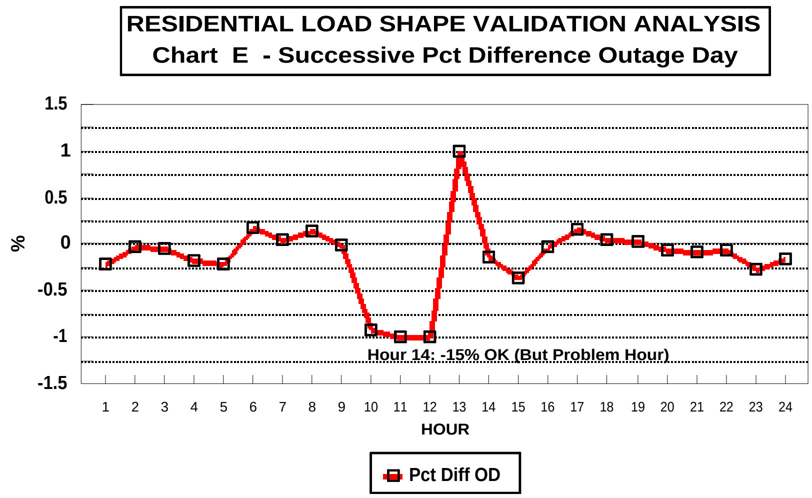


Successive Percentage Change

It is logical to assume that an outage such as that portrayed here should be identifiable by checking the change in load, which should be substantial if the loads dropped to zero, then spiked after the outage. This test has often been applied by load analysts in the past.

Applying the technique of successive percentage change on the right (Chart D) to the typical day (Chart A) results in as much as a 70% change. You can see how variant the test results can be. A legitimate (and common) increase in load due to morning activity produces a very large percentage increase that would be flagged as excessive.

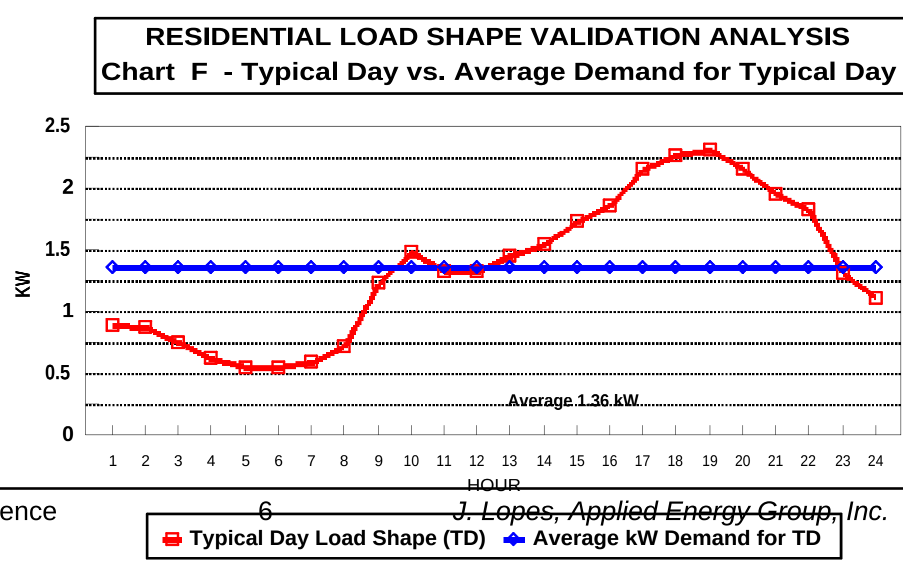




A similar test applied to the outage day shown above in Chart B produces Chart E at right. While the largest changes are several 1's (100%), corresponding to the first full outage hour and first payback period, applying a criteria would be difficult and arbitrary. A 90% change threshold for rejecting the interval would flag hours 10 through 13, obvious problem hours, but would <u>not</u> flag hour 14, since it did not change much compared to the previous hour, which was invalid due to continued payback. Thus, the successive percentage validation test did <u>not</u> produce the desired objective of flagging all the most serious problems and since it missed hour 14's payback spike. Overall, the advantages of this test are its simplicity and ability to identify when an outage starts. It's disadvantages include the difficulty in setting a threshold value to reject, since even typical days have significant changes from one interval to the next. It is also a biased test, since some successive intervals are naturally increasing, like in the mornings, and some naturally decreasing, like in the evenings. If a 50% increase is common for hours 9 - 10 and an outage caused a 50% decrease, there would be no way to set a criteria to flag that.

<u>Pattern Recognition</u>

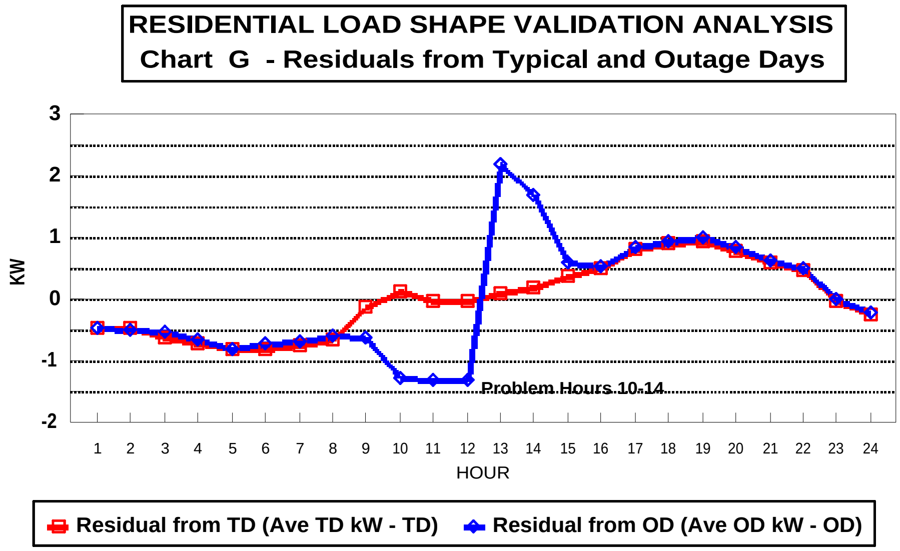
In this type of test, a base load pattern is constructed, then applied interval by interval against each day's load shape to identify intervals that deviate too much from the base pattern. This concept holds much promise, but the key is in how you define the pattern. Ideally, a pattern would be independent of load level, so that a base pattern could be scaled up to match load levels of each day and so that interval outages could be distinguished easily.



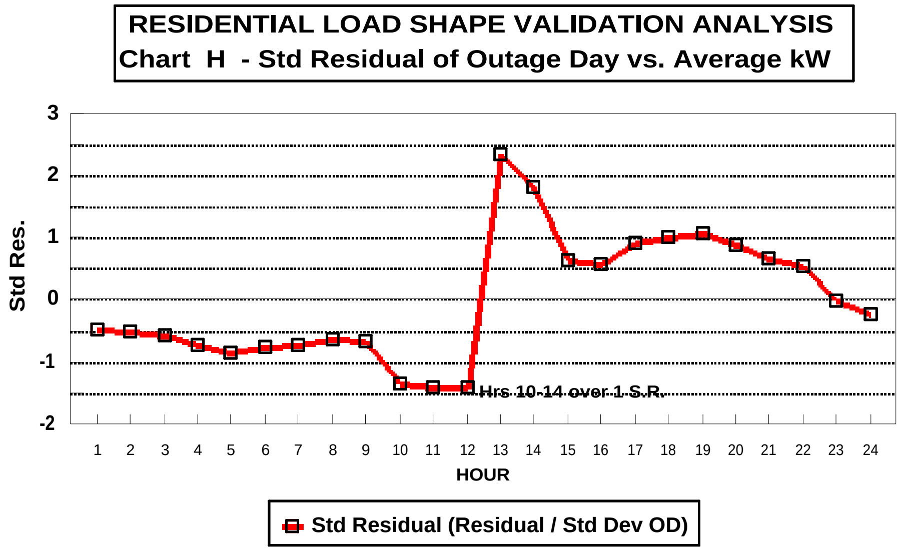
Based on many years of processing of load data, a successful pattern recognition algorithm observed is to define a residual, calculated as the difference between each interval's value and the average kW demand for the day. As indicated in Chart F, the average value of 1.36

kW is compared to each interval.

The residual pattern is calculated as the difference between each interval value and the average daily value. The deviation (residual) of each interval from the average daily kW value is first computed (shown on Chart G for both a typical and outage day)



The set of residuals is then divided by the standard deviation of the daily interval values to produce a standardized pattern. Chart H shows the standard residuals for the outage day.



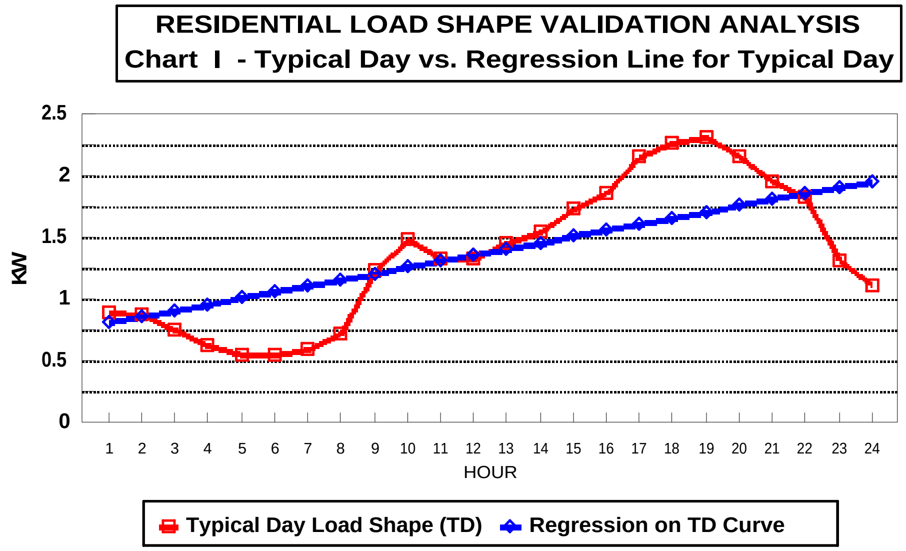
A base pattern is calculated for each month or season (as necessary) for each day type (Weekday, Saturday and Sunday/Holiday).

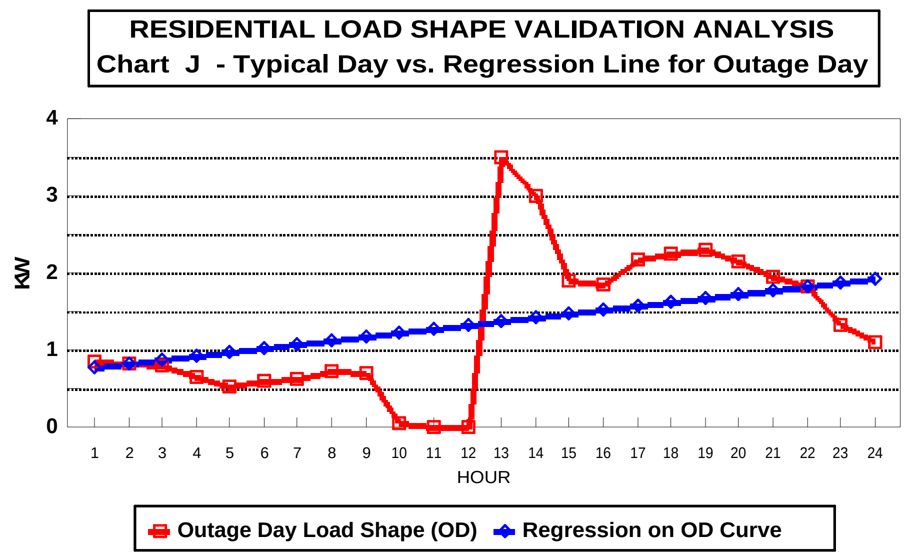
The user must decide on a tolerance for the pattern test (in number of standard deviations). A higher tolerance will mean that fewer intervals will fail the tests. The tolerance level should be based on the percent of intervals that the user wishes to have fail to ensure that real problems are flagged. When comparing the new day of data to the historical interval pattern, the same procedure of computing standardized residuals is performed on the current day. The current day's standard residuals are then compared to the base interval pattern (for the same data track) to see whether it deviates from the historical pattern by more than the allowed tolerance.

In the example above (Chart H), the outage occurs during hours 10, 11 and 12, then the payback spike during hours 13 and 14. Based on the results above, it is logical for the criteria to be set at +/- 1 standard residuals, since every one of the problem hours (10-14) exceeds one standard residual from the average, and none of the “good” intervals exceeds one standard residual.

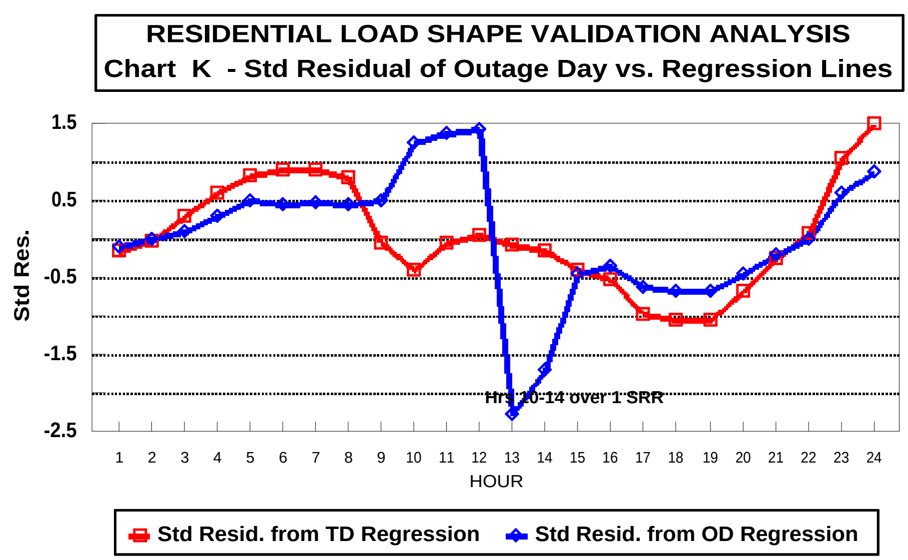
This test obviously has advantages, including being unbiased and applicable even where load levels may be different. So, for example, a residual pattern baseline from a previous month can be used for the next month, even if weather is somewhat different. Data tracks can be validated against their own pattern from a recently validated set of load data. Both spikes and troughs will be flagged, along with any subsequent intervals that have not returned to normal shape. In addition, if there is no historical data, an average load shape can be developed for each site and used to construct a baseline upon which the same period’s data can be checked for each individual day. The disadvantage is that it cannot identify when all the intervals are missing, whether zero intervals exist (these could easily pass this test) or all intervals are a multiple factor higher or lower than they should be. This test should therefore be combined with other tests that can identify invalid load levels, such as high-low daily consumption tests by day type, or simple checks for zeros that would be obvious flags regardless of other problems.

An improvement upon this approach is to use a regression line, rather than the average kW demand line, to construct the base for the residual calculations. As indicated in Chart I for a typical day and Chart J for an outage day, the regression line reflects the increasing load during the day, so will produce a more unbiased and stable baseline from which to calculate the standardized residuals. This also helps when weather changes within a day affect the load shape.





In Chart K, the standardized residuals for the regression-based base pattern clearly show that a criteria whereby standard residuals greater than 1 would indicate problem data and would accurately flag hours 10 through 14 as the only problems for the outage day. They would only indicate hour 24 and possibly 23 for the typical day, and this would be a Type I error, which is acceptable under the stated objectives of the validation testing. Thus, pattern recognition, when designed properly, can be a very effective method of validation testing.



EDITING RULES

Once the problem intervals are identified, decisions on how or whether to patch must be made. Depending on the extent of the problem data that required patching, different approaches could be used.

Smoothing — For short-term outages of, for example, four intervals or up to 1-4 hours, valid values on either side of the problem period would be averaged and the result inserted into each problem interval. This would not be advisable for longer periods, since it essentially inserts a flat load, which is not necessarily applicable to a long time period.

Shaping — For short-to-medium periods, typical shapes for the customer from a library of that customer's similar day types can be scaled to the load levels for the missing periods and applied to part of a day, or to full days if they can be adequately categorized.

Borrowing — For longer problem periods, borrowing data from a similar day can be used. Typically, weather, day type (weekday, Saturday, Sunday) or specific day of the week is matched for the same customer and used as the source for borrowing data for periods of less than a day to several weeks. This preserves the appropriate shape, but the number of conditions means that preparation must be made to ensure that it can be used for automated routines.

Exclusion — When too much data must be patched, it may become risky to "manufacture" a load shape if it is not necessary. In cases where sample sizes are sufficient to justify settling for fewer sample points, this option is reasonable. This may not be an option for larger customers for which the site does not represent a sample of a group, but rather its own group. In these cases, however, the loads are typically less variant and borrowing of data from an earlier period or averages of load patterns is much less risky than for smaller customers.

Patterning — Similar to shaping, this technique uses the standardized residual pattern to patch short time periods of missing or invalid data. There must be sufficient valid intervals to establish the use level, so it is more applicable to outage days than to days where equipment malfunctions cause missing intervals.

In terms of typical examples of problem data and the technique considered most applicable under the most stringent of dynamic load profiling conditions, the following are suggested:

Missing less than four intervals - Smooth missing data period

Outage involving less than four intervals - Smooth outage and payback period identified by pattern recognition algorithm

Missing or outage period of between four intervals and a half day - Patterning using base patterns from the same day of the week and comparable weather. These can be set up as variables in advance.

Missing or outage whole day - Borrow data from same day of the week either one or two weeks before. As an alternate, same day type (weekday, Saturday, Sunday). Fridays or Mondays may be distinct.

Overall, a critical factor in editing is consistency, whether by automated algorithms where it can be programmed, or semi-automatic or manual processing.

EDITING EXAMPLES

Chart L illustrates the corrected intervals applied to the Outage Day originally presented in Chart B, overlaid on the original Outage Day shape. Hours 10 - 14 have been replaced by the values obtained from converting the base pattern for the day type to levels comparable to the overall use level for the specific day.

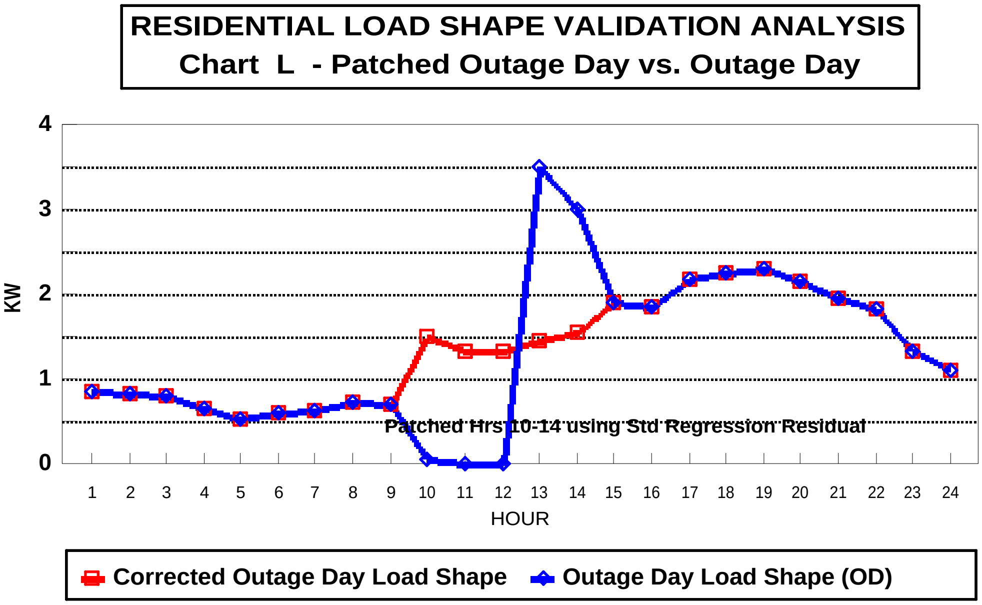
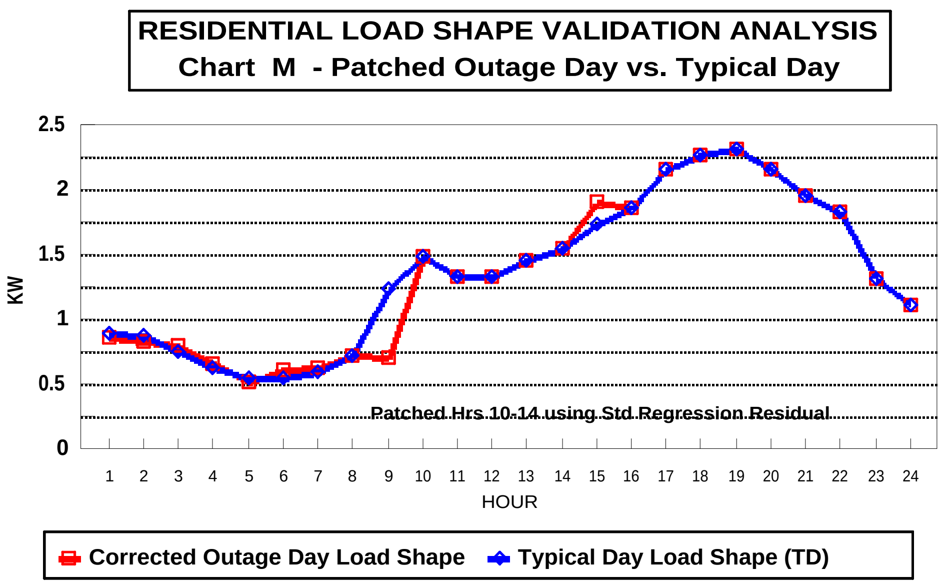


Chart M compares the corrected Outage Day to the Typical Day presumed if there had not been an outage. Note that hours 9 and 15 were not corrected because they were not sufficiently different from the base pattern. The outage was in effect during only part of hour 9 and the payback was almost completely resolved by hour 15.



In any event, any patching or changes to raw load data should be documented through an audit trail to ensure that too much editing is not being done, and a means to restore the original version should be in place to maintain the billing parameters or to reverse the editing, should that be considered necessary later on.

HOW HAS THE DEREGULATED INDUSTRY CHANGED THE WAY WE LOOK AT VALIDATION AND EDITING?

While it is one thing to depend on visual inspection of the data, supplemented by graphic or diagnostic analysis of the load intervals, the increasing need for real-time data for support of real-time load expansions to drive load profiling models under deregulation's market settlement process, means that a more automated technique must also be available. In the case of dynamic load profiling, as applied in California and some other states, load data from sites is remotely read, processed and feeds a model that expands all load sites to form class, supplier and system load profile estimates "on the fly". During this process, if there is problem data, there is often no option for analyst intervention to "leisurely" review the data for, particularly, the two greatest problems - recorder malfunctions and real outages. Recorder malfunctions, particularly data communications failures, may be detectable by the systems used to collect the data, so that may solve the problem. However, the detection of real outages, whereby a drop and payback spike is accurately recorded, requires analytical techniques, since there is no actual failure involved. An automated validation algorithm is the only way to realistically address this when the number of sites and the time available to edit becomes too large to reasonably expect human intervention to keep up. Even with problem data identified, the decision on whether to patch the problem data cut or drop the whole time period from the analysis must be made. While it might be easier to drop any problem cuts, the sample sizes may not be able to withstand the number of drops, which would reduce the sample size - and precision - or else the site may be large and only represent itself, so dropping it would reduce the load of the group it is a contributor to, and seriously skew the results.

With the automated pattern recognition validation techniques presented here, along with automated editing using the base pattern and specific editing rules, dynamic profiling can be accomplished with less risk. Given more time, more comprehensive validation and editing techniques can be applied, as well as old fashioned eyeball inspection, but the techniques for dynamic profiling <u>must</u> minimize risk as their primary priority.

SUMMARY

Traditional validation and editing techniques have a place in the new deregulated environment, but they must be applied selectively and effectively to reduce the risk associated with being wrong - either by missing problem data that could seriously skew results, or by too much averaging and smoothing. Techniques that cannot translate to computer-based automated routines may not be useful for the speed of real-time load profile expansion. There must be an automated solution to every conceivable problem. The techniques presented in this paper have been applied by the author in load research projects over the last 23 years and have proven their validity and value. I can only hope that they have inspired thought and consideration as the readers march into the frontiers of deregulation.